THE POSSIBILITIES ARE INFINITE FUJITSU

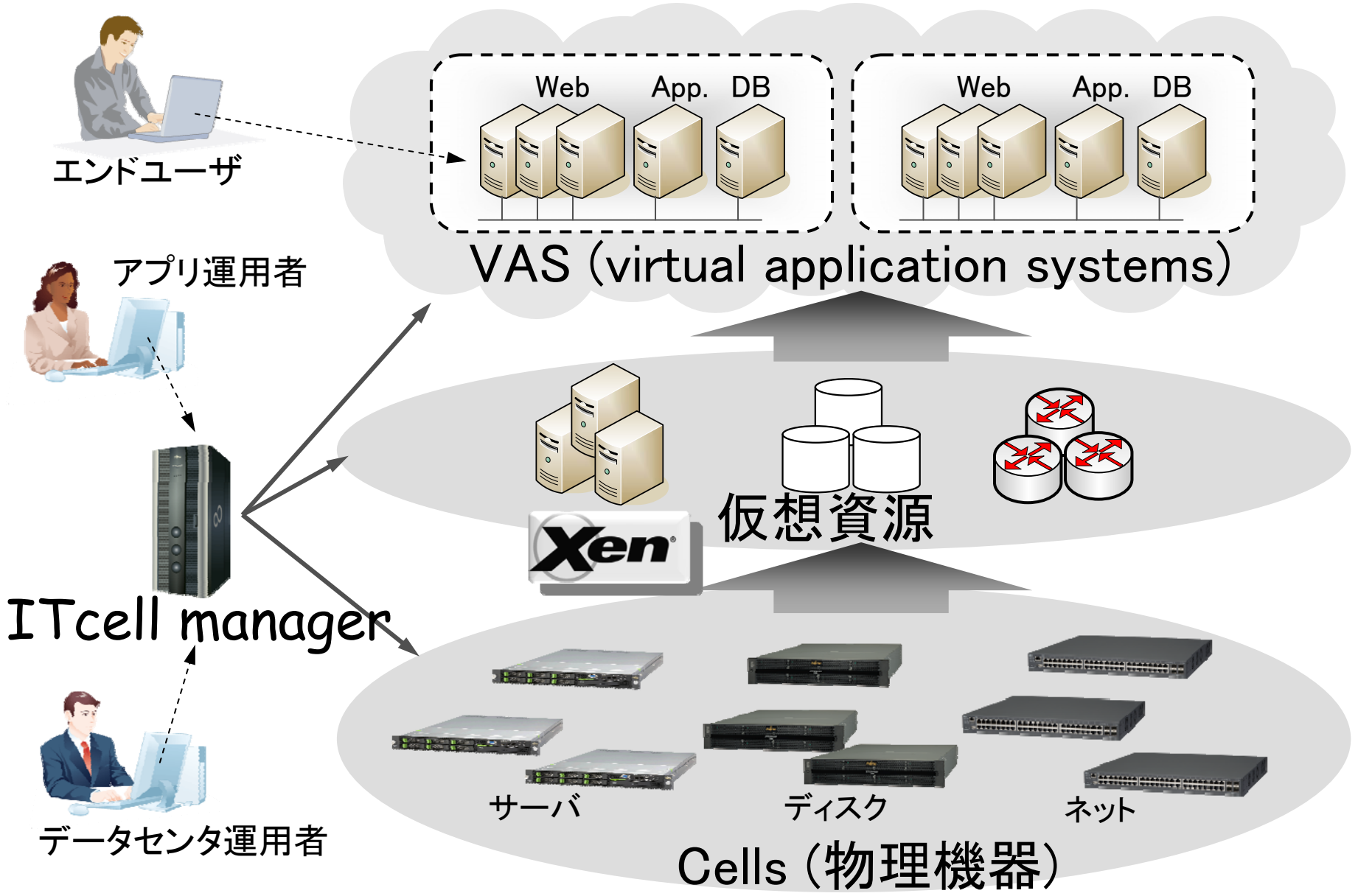
クラウド、作ってみた。

IaaSのネットワーク設計・運用・監視の実際

今井祐二
株式会社富士通研究所
クラウドコンピューティング研究センター

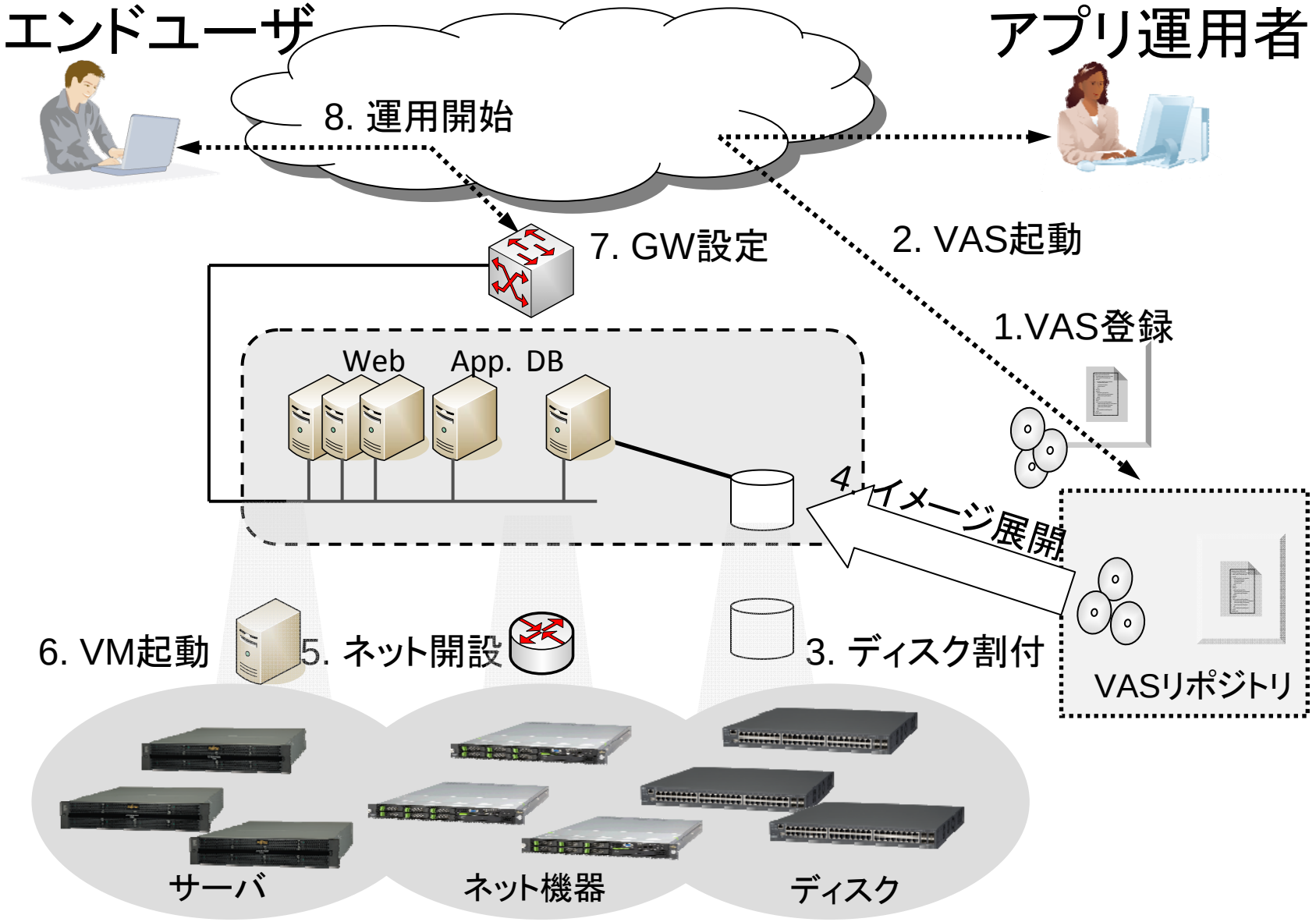
IT cell（富士通研IaaSプロトタイプ）

FUJITSU



仮想アプリシステム起動の流れ

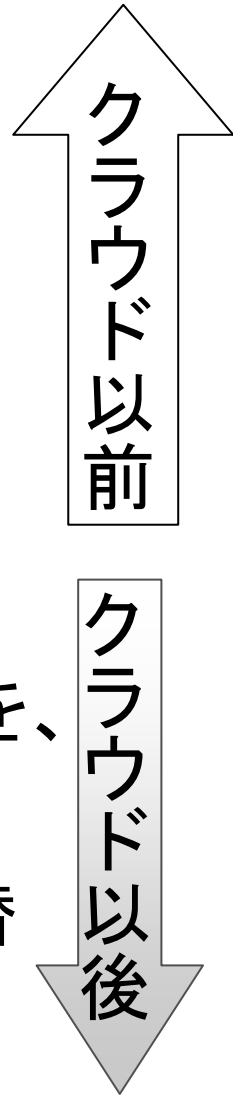
FUJITSU



データセンタのハード構成単位

FUJITSU

- ホスティング
  - マルチテナント可能なアプリケーションで大規模ユーザを収容
- ハウジング
  - 顧客ごとのオーダーメードシステムをお預かり
- ハードウェアプール
  - 共通プラットフォームとなるサーバ・ネットワークを、可能な限り大きなロットで設置
  - 中長期需要見通しに基づいて、ロット単位に入替



事前の構想

FUJITSU

- 顧客仮想システムは分離して運用
  - Xenを使ってOSレベルで分離　…*OK!*
  - ストレージはiSCSIをdom-0経由でxsd(仮想ブロックデバイス)として見せる。　…*OK!*
  - ネットワークはVLANで切る。…*工夫してみました。*
- それなりに冗長性作り込みもしてみよう。
  - ストレージはRAID-1、L2はdom-0でeth0, 1をbonding、スイッチも2重化　…*OK!*
  - 冗長化された部分が故障したら、別ノードにVMマイグレーション …*OK!*

紹介する工夫ポイント

FUJITSU

1. VM間IPトンネル
   - 仮想システム間ネットワーク分離
2. Rack & Go
   - ネットワークによる物理インストレーション支援

1. VM間IPトンネル
仮想システム間ネットワーク分離

VLAN設計の選択肢

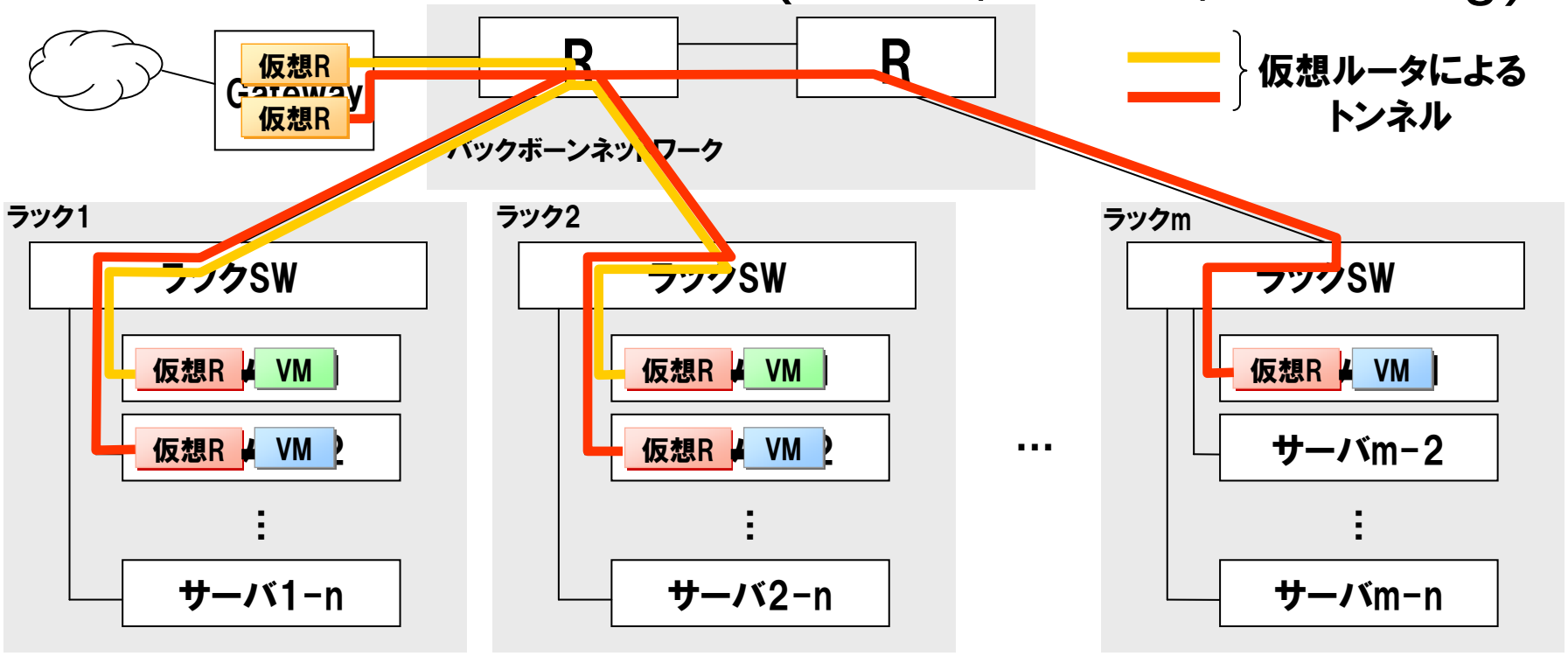
FUJITSU

1. Vanilla-L2ネット
   - すべてのtagフレームを素通し。ネット隔離はdom-0の仮想スイッチが、どのtagで足を出すかで制御
   - STP系の使用を避けると、ループフリーなトポロジにしておきたくなるが、スケーラビリティに限界が出そう
2. VM配置に連動してtag-VLANを切る
   - 仮想システムの起動・停止・VM増設・マイグレーションで、猫の目のようにスイッチの設定が変化
   - 正しく設定できるか?正当性を検証・監査できるか?
   - 不測の事態に対応できるエンジニア確保が大変そう

工夫してみました。(VM間IPトンネル)

FUJITSU

- ユーザVMと1:1に仮想ルータ用VMを起動。VM間トラフィックをIPトンネルでくるむ。
- 外部接続用GWにも仮想ルータVMを起動。トンネルと外のVLANを中継。
- バックボーンはプレーンL3 (OSPF, VRRP, bonding)



VM間IPトンネル　利点・欠点・がんばりどころ

FUJITSU

- 利点
  - L3バックボーンは「いつものやつ」
    - ネットワークエンジニア・オペレータは容易に理解
    - 障害時迂回経路・ネット拡張はOSPFにおまかせ
    - ブロードキャスト問題から開放
  - ネット分離検査は、トンネル&経路設定チェックでOK
- 欠点
  - 見慣れない組合わせ。「VM」と「IPトンネル」
    - アプリエンジニア・オペレータへの説明が大変
- 評価中/これからのがんばりどころ
  - 仮想ルータによる性能劣化は許容範囲か?

2. Rack & Go

ネットワークからの物理インストレーション支援

日々是インストール

FUJITSU

- 想定

  150万台物理サーバプールを3年償却で保持

- 年50万台のサーバ入替（≒2500台/日）
  - ラックへのマウント
  - ケーブリング(ラック内・ラック間)
  - インストレーション(BIOS設定・OS・ミドルウェア)
  - 検査
  - 故障交換時にはノード単位で同様の作業が発生
  - どうしろっていうんだ・・・。

工夫してみました。(Rack&Go)

FUJITSU

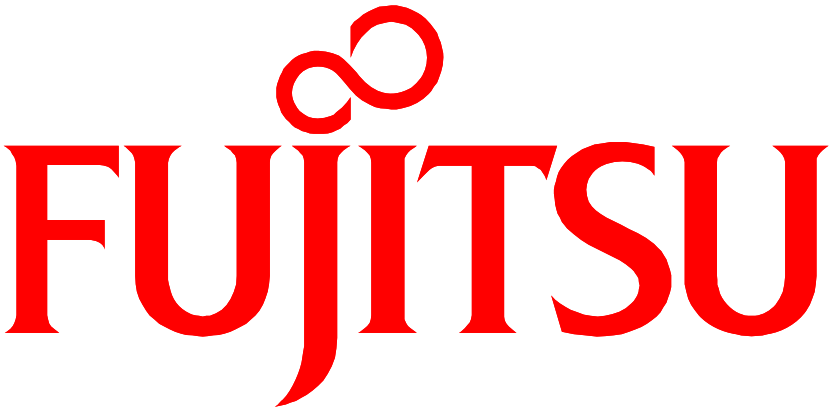
- サーバ物理ラッキング終了時に、USBメモリを刺して電源を入れて待っているとインストールが完了

1) サーバ搭載・ネット結線情報をDB登録
2) 工場で事前ラック搭載、データセンタ搬入、設置＆ケーブリング
3) Fedora Live USBメモリを挿入し、電源ON
   - テンポラリIPアドレスをDHCPで獲得
   - IPMI SOL (Serial over LAN)をenable
   - MACアドレスをサーバに報告(ipmi, eth0, eth1)
4) SWのMAC学習テーブルをスキャン
   - サーバの各I/F MACが正しいポートに出てくるか検査
5) dhcpdに正式なアドレス(IP & MAC)登録
6) IPMI SOL経由でBIOS設定(ipmitool & expect)
7) PXE Bootでソフトウェアインストレーション
   プロビジョニングOSS各種(Kickstart, Cobber, Puppet)

Rack & Go　利点・がんばりどころ

FUJITSU

- 利点
  - インストレーション品質向上は実感
    - ・ケーブル結線異常は確実に検出
    - ・上位層担当SEの物理設定エラーへの懸念解消
- これからのがんばりどころ
  - 作業効率向上は本当にあがったのか評価が必要
  - サーバBIOS設定の機種依存部を吸収したい
  - netconfでのネットワーク機器設定もやりたい



THE POSSIBILITIES ARE INFINITE